Code 1

THE TRANSFORMATION OF SENTENCES FOR INFORMATION RETRIEVAL

Jane J. Robinson

December 1965

P-3243

THE TRANSFORMATION OF SENTENCES FOR INFORMATION RETRIEVAL

Jane J. Robinson*

The RAND Corporation, Santa Monica, California

The sentence as a unit of language stands midway between the word and the paragraph. If words are the basic units for classification and indexing and paragraphs the basic units for abstracting, the sentence is the basic unit for fact retrieval.† Very simply, the central problem of fact retrieval is: Given an interrogative sentence, how does one recognize a matching sentence that supplies an answer? The simplest case is a sentence beginning with an interrogative word followed by a string of additional words, matched by a sentence which replaces the interrogative with an answering word or phrase.

Who	invented the flying shuttle?
John Kay	invented the flying shuttle.

---

*Any views expressed in this paper are those of the author. They should not be interpreted as reflecting the views of The RAND Corporation or the official opinion or policy of any of its governmental or private research sponsors. Papers are reproduced by The RAND Corporation as a courtesy to members of its staff.

This paper was presented at the 1965 Congress of International Federation for Documentation (FID), Washington, D.C., October 1965.

†"Fact retrieval" is not a well-defined term; "data retrieval" or "text retrieval" are substitutes. All that is intended here is to distinguish between the problem of providing references ("document retrieval") and the problem of providing statements within documents that can answer specific questions. The problem of truth is something else again and lies outside the scope of syntactic analysis as treated in this paper.

If all cases were so simple, a computer could be programmed to find the matching sentences within a large store of text much more easily, accurately, and completely than humans could. As usual, the simplest cases are vanishingly rare, and so far no computer programs can cope adequately with the shifting word orders and the forms, sometimes protean, sometimes elliptic, that sentences in natural language texts most frequently take.

The difficulty is that the basic meanings represented by sentences are not isomorphic with their surface forms, and the computer can deal directly only with forms. In terms of meaning, the sentence

John Kay invented the flying shuttle in 1733.

is the matching answer for both

By whom was the flying shuttle invented?

and

When was the flying shuttle invented?

But these examples have already complicated the mechanical definition of the procedures for recognizing the match. It is more complicated still to provide for mechanical recognition of matches within sentence boundaries, where the answer is contained in phrases such as: ". . . the inventor of the flying shuttle, Kay . . ."; ". . . Kay's invention of the flying shuttle in 1733 . . .," etc.

I have posed the problem in terms of finding a mechanical procedure for question-answering, not primarily to assess the state of the art of automatic language processing for information retrieval, but because these terms

make more clear and concrete the general problem of recognizing relevance and sameness of meaning at the sentence level in spite of formal differences. (The related problems of synonymy at the word level and of pronoun reference across sentence boundaries may be more amenable to solution if the problems of sentence structure are solved first.) Heuristic methods for dealing with any single paradigmatic set of examples of the sort cited above are possible, but heuristic or ad hoc procedures have, so far, proved inadequate to deal with the bewildering variety of sentences in natural text. We need a general procedure firmly grounded on an understanding of the basic processes of sentence construction provided by the grammar of a language. We cannot tell a computer how to recognize paraphrases unless we understand how we ourselves recognize them.

Of course, sameness of meaning and difference of form confront us all the time. Our universes of experience and of discourse are both in a constant state of flux and no man ever steps into the same river or says the same thing twice. The river, the acoustics, and the man change through time. Yet all our acquisition and organization of knowledge rests on our perceiving similarities and continuities, in spite of objective differences. For various human purposes, we regard different items of experience as instances of the same thing and we judge their differences to be irrelevant. Moreover, we can communicate our knowledge to each other with ease and accuracy only to the extent that we ourselves are similar. Only through shared experience and shared conventions can we speak the same language and classify documents in the same terms.

But if communication implies a community of custom and of language, various layers in that community can tolerate varying amounts of divergence from convention. So long as individual behavior does not deviate from some basic set of implicitly defined conventions, eccentricities and idiosyncrasies are tolerable. In some areas of behavior, however, we must eliminate individual differences in the interests of communication. A detailed account of a laboratory experiment is written for the most part in the passive voice: the centrifuge was operated and the amount of the isotope was measured, and the individual characteristics of the operator and the measurer should not matter. The terms of the scientist are more rigorously defined and his sentences more conventionally constrained because his statements and descriptions often presuppose interchangeability among observers and experimenters. Thus, attempts to automate translation from one language to another have started with scientific reports rather than with poetry.

So also, the amount of tolerable divergence differs from layer to layer within language. If we are to communicate at all, the conventions of language are most sharply defined and restrictive at the lowest level--that of the basic units, the phonemes, and the letters. New words come easily into our vocabularies, but the phonemes that represent them, the letters that in turn represent the phonemes, and the rules for combining them into syllables, change with glacial slowness. At a higher level, the vocabulary appears to be not only larger, when we compare the stock of morphemes or words to the stock of

phonemes, but subject to more rapid change. It is a relatively open system. But the new words are for the most part nouns and verbs like astrogation, astrogator, and astrogate, whose parts are familiar. Furthermore, the most frequent words of our vocabulary--the pronouns, the prepositions, the auxiliaries, etc.--show little alteration through time.

The number of letters is finite and small; the number of words or morphemes is finite though large. Given an alphabet, therefore, a computer can match letters and words with mechanical regularity, and relieve us of the work of making indexes and concordances. But when one comes to the level of the sentence, the possibilities are infinite. Setting aside those instances of quotation and barring multiple copies of the same document, how many times can one expect to find a repretition of any given sentence in a large collection of documents? If "sentence" is defined as any stretch of words between one mark of end punctuation and another, the probability of finding a repetition is extremely slight.

The reason for this flowering of individuality at the sentence level, the property of natural languages that both provides for it and makes it tolerable to the community, has become clearer in recent years, principally through the theoretical work in linguistics primarily associated with Chomsky and Harris and their respective schools [1,2,3]. Briefly, it is because the rules for sentence construction are recursive; that is, a basic sentence unit or "kernel" can embed within itself another basic unit, which can embed another in turn, and so on ad infinitum. Some embeddings

are obvious at the surface, as in

> Lewis Paul knew that John Kay invented the flying shuttle.

More often they are transformed, as in

> Lewis Paul knew John Kay, the inventor of the flying shuttle,

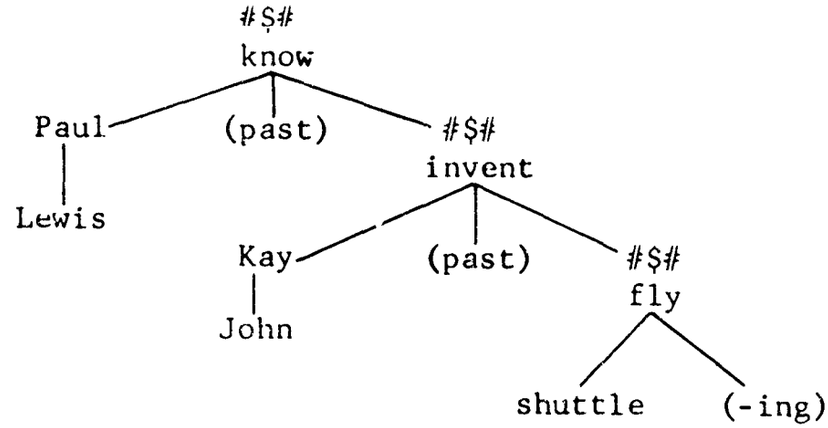
or

> Lewis Paul knew about the invention of the flying shuttle by John Kay.

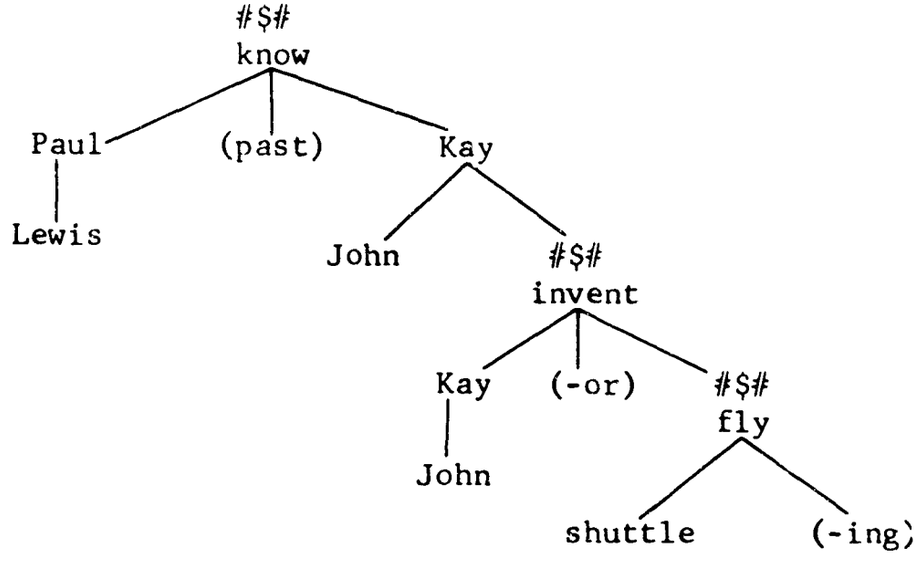
The dependency graphs [4] of Fig. 1 show how the underlying, untransformed, basic structures embedded in these three sentences might reasonably be represented.

These graphs exemplify the reduction of different surface structures with the same basic meaning to strongly similar, embedded, "canonical" forms representing that meaning. Such a reduction, a many-one mapping of surface structures onto a relatively few deep structures, suggests a finite "alphabet" for sentences, roughly analogous to the alphabet for words, so that mechanical matching procedures for meanings through the matching of forms can become feasible. Even if mechanical procedures prove impracticable, the insights gained into the representation of meaning, especially the representation of the "same" meaning in formally different sentence structures, may help us devise more standardized ways of storing information and constructing data bases for question-answering or deductive systems in information retrieval.

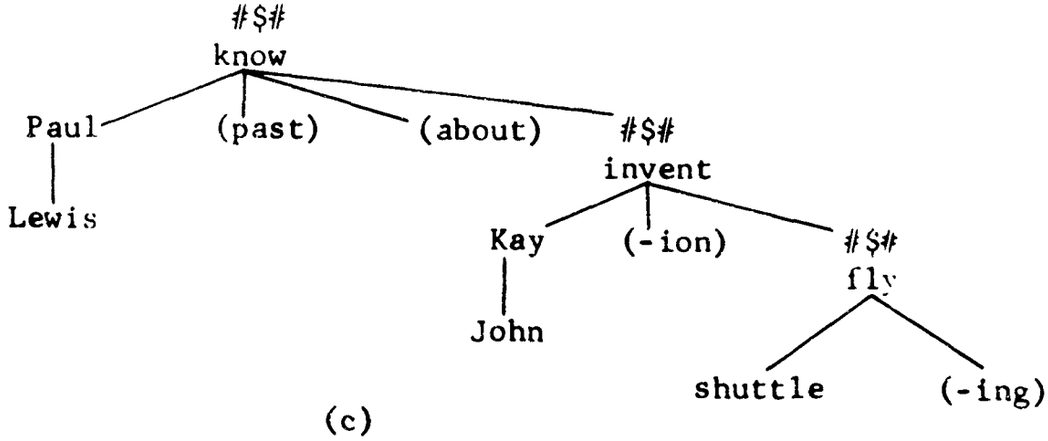
It is not the sentences, but their kernels that appear to be the units for representing meaning. One important


#$#
know
Paul
(past)
#$#
invent
Lewis
Kay
(past)
#$#
fly
John
shuttle
(-ing)

(a)


#$#
know
Paul
(past)
Kay
Lewis
John
#$#
invent
Kay
(-or)
#$#
fly
John
shuttle
(-ing)

(b)


#$#
know
Paul
(past)
(about)
#$#
invent
Lewis
Kay
(-ion)
#$#
fly
John
shuttle
(-ing)

(c)

Figure 1

point is that embedding and transformation permit the construction of sentences containing many basic meanings, related to each other in various ways, and this in a sense makes sentences more efficient storage devices; consequently a single sentence often provides answers to many different questions. Also, the point of view of the questioner need not be strongly similar to that of the writer whose sentence contains an answer. The surface structure of the writer's sentence can reflect some of his immediate purposes for organizing his information, for emphasizing some aspect of it and subordinating others, as well as his individualities of style. The structure of the questioner's interrogative can reflect a different immediate purpose and a different style. Communication is still possible, because the deep structures of their sentences adhere to the same conventions.

In the last six years, several research groups have attacked the problem of designing automated question-answering systems based on natural text rather than on highly structured data bases, and various techniques for combining syntactic and semantic analyses have been used [5,6,7]. The view adopted here is that semantic (and other) techniques will prove more effective if applied after a syntactic analysis that explicates the deep structures. That, of course, depends upon the development of detailed transformational grammars.

This view is borne out by the difficulty encountered by current automated parsing grammars assigning structural descriptions directly to sentences. Applied heuristically, they miss valid structural assignments that correctly

correlate an expression with equivalent paraphrases and relevant questions [8,9]. Applied algorithmically, they produce an unmanageable number of parsings, and a surprising proportion of them correspond to possible ambiguities and thus are not eliminable. Many, if not most, of these ambiguities arise because the transformation of embedded sentences may lead to constructional homonymity on the surface of the sentence, as in the famous "Flying planes can be dangerous." Moreover, the necessary co-occurrence rules become unmanageably numerous if written for all surface structures rather than for the smaller set of deep structures.

One avenue to be explored is to subject each of these multiple analyses of a sentence produced by a loosely constructed "surface" grammar to inverse transformations, comparing the results with a tightly constructed grammar to find the simpler deep structures from which any valid surface structure must be derived. For example, the surface grammar would, typically, produce two analyses for "John was drunk by midnight": one would label it a passive, corresponding to "Midnight drank John." Comparison of this inversely transformed kernel with the requirements of a precise deep grammar, however, should reveal the presence of co-occurrence restrictions on inanimate "time" nouns with verbs like "eat" and "drink," which require animate subjects.

The major linguistic task, then, is to provide detailed, analytic, recognition grammars with transformational components adequate to deal with the complexities of the surface structures of natural sentences if the necessarily ad hoc but ultimately unsatisfactory simplifying assumptions

of current question-answering systems are to be superseded. Until quite recently, transformational grammars have been written to generate rather than to analyze, although as early as 1961 Matthews [10] proposed a technique for analyzing a given sentence by synthesis from a generative grammar.

Work on the recognition problem is now underway, and three different types of grammar are evolving with transformational components designed to recover deep structures automatically. Kuno [11] reports some experiments with the Harvard predictive analyzer to produce kernel sentences concurrently with the analysis of surface structures. Petrick [12], Kay, and the MITRE Language Processing Techniques Subdepartment [13] have all proposed methods applicable to phrase structure grammars. An "approximate" formalism to obtain structural descriptions similar to deep structures is being developed by Lieberman, et al., at IBM [14]. Although applied to a phrase structure grammar now, the formalism is intended to be applicable to other models as well. Robinson experimented briefly with a paraphrasing routine for a phrase structure grammar [9], but is currently designing a dependency grammar with transformations, in collaboration with Hays and Kay.

Several machine translation groups are also incorporating transformational features into their grammars, in accord with Harris' assumption that many languages are more similar in their kernel sentences than in their total surface structure. Linguistic work in translation is obviously an important part of information retrieval, but can only be mentioned here.

It would be unrealistic to suppose that practical programs for automating the retrieval of information expressed in natural text will be forthcoming in the next few years. Experience with machine translation has shown that to extrapolate from progress made in early stages with simpler patterns of natural languages can lead all too easily to speciously optimistic predictions of early success. Nevertheless, a cautious optimism can be based upon certain signs. Detailed knowledge about the languages is accumulating. At the same time, the capacity of computers to handle masses of non-numerical information is increasing and the iteration between man and machines is becoming easier as well as faster. Most promising of all, from a linguist's point of view, is the development of a theoretical framework in linguistics within which can be fitted the description of the covariance of form and meaning at the syntactic level, extending beyond the morpheme and the word and into the sentence, where propositions are stated and interrogated.

REFERENCES

1. Chomsky, N., Aspects of the Theory of Syntax, The Massachusetts Institute of Technology Press, Cambridge, Massachusetts, 1965.

2. Fodor, J. A., and J. Katz (eds.), The Structure of Language: Readings in the Philosophy of Language, Prentice-Hall, Inc., Englewood Cliffs, 1964.

3. Katz, J., and P. M. Postal, An Integrated Theory of Linguistic Description, Research Monograph No. 26, The Massachusetts Institute of Technology Press, Cambridge, Massachusetts, 1964.

4. Hays, D. G., "Dependency Theory: A Formalism and Some Observations," Language, Vol. 40, No. 4, October 1964, pp 511-525.

5. Hays, D. G., and R. Ma, Computational Linguistics: Bibliography, The RAND Corporation, RM-4523-PR, March 1965.

6. Simmons, R. F., Answering English Questions by Computer: A Survey, System Development Corporation, Santa Monica, California, SP-1556, April 2, 1964.

7. Salton, G., "Automatic Phrase Matching," Preprint of a paper presented at the 1965 International Conference on Computational Linguistics, New York, May 1965.

8. Kuno, S., and A. Oettinger, "Syntactic Structure and Ambiguity of English," AFIPS Conference Proceedings, Vol. 24, 1963 Fall Joint Computer Conference, pp. 397-418.

9. Robinson, J., The Automatic Recognition of Phrase Structure and Paraphrase, The RAND Corporation, RM-4005-PR (Abr.), December 1964.

10. Matthews, G. H., "Analysis by Synthesis of Sentences in a Natural Language," 1961 International Conference on Machine Translation and Applied Language Analysis, Her Majesty's Stationery Office, London, 1962.

11. Kuno, S., "A System for Transformational Analysis," Preprint of a paper presented at the 1965 International Conference on Computational Linguistics, New York, May 1965.

12. Petrick, S. R., "A Recognition Procedure for Transformational Grammar," presented at the 2nd Congress in the Information System Sciences, The MITRE Corporation, Bedford, Massachusetts, November 1964.

13. Walker, D. E. (ed.), English Preprocessor Manual, Information System Language Studies Number Seven, Language Processing Techniques Subdepartment, Information Sciences Department, The MITRE Corporation, Bedford, Massachusetts, SR-132, December 1964.

14. Lieberman, D., D. Lochak, and K. Ochel, "Automatic Deep Structure Analysis Using an Approximate Formalism," Preprint of a paper presented at the 1965 International Conference on Computational Linguistics, New York, May 1965.